Two-Stream Spectral-Temporal Denoising Network for End-to-End Robust EEG-Based Emotion Recognition

Xuan-Hao Liu[1], Wei-Bang Jiang[1], Wei-Long Zheng[1], and Bao-Liang Lu[1,2,3(✉)]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China
{haogram_sjtu,935963004,weilong,bllu}@sjtu.edu.cn
[2] RuiJin-Mihoyo Laboratory, RuiJin Hospital, Shanghai Jiao Tong University School of Medicine, 197 Ruijin 2nd Road, Shanghai 200020, China
[3] Key Laboratory of Shanghai Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China

Abstract. Emotion recognition based on electroencephalography (EEG) is attracting more and more interest in affective computing. Previous studies have predominantly relied on manually extracted features from EEG signals. It remains largely unexplored in the utilization of raw EEG signals, which contain more temporal information but present a significant challenge due to their abundance of redundant data and susceptibility to contamination from other physiological signals, such as electrooculography (EOG) and electromyography (EMG). To cope with the high dimensionality and noise interference in end-to-end EEG-based emotion recognition tasks, we introduce a Two-Stream Spectral-Temporal Denoising Network (TS-STDN) which takes into account the spectral and temporal aspects of EEG signals. Moreover, two U-net modules are adopted to reconstruct clean EEG signals in both spectral and temporal domains while extracting discriminative features from noisy data for classifying emotions. Extensive experiments are conducted on two public datasets, SEED and SEED-IV, with the original EEG signals and the noisy EEG signals contaminated by EMG signals. Compared to the baselines, our TS-STDN model exhibits a notable improvement in accuracy, demonstrating an increase of 6% and 8% on the clean data and 11% and 10% on the noisy data, which shows the robustness of the model.

Keywords: EEG · EMG · Emotion Recognition · End-to-end · Denoising · Robust Classification

1 Introduction

The rapid development of deep learning techniques opens up new possibilities for brain-computer interfaces (BCI). Different from the BCI applications in helping

B. Luo et al. (Eds.): ICONIP 2023, LNCS 14449, pp. 186–197, 2024.
https://doi.org/10.1007/978-981-99-8067-3_14

a paralytic patient walk again [1], affective brain-computer interface (aBCI) is aiming to detect, analyze, and respond to human emotions. Emotions are essential in our daily lives and influence our behaviors and mental states consciously or unconsciously.

In recent decades, EEG-based emotion recognition has obtained great interest due to the reason that EEG signals are inherently correlated to brain activity [2]. The previous EEG-based emotion recognition studies are almost based on manually extracted features, e.g., power spectral density (PSD) [10] and differential entropy (DE) features [14]. Significant progress has been achieved by using these handcrafted features [10,17]. However, they could be biased in specific domains and ignore rich information in the temporal domain. To fully excavate emotion-related information in raw EEG signals and eliminate the complicated process of handcrafted feature extraction, end-to-end models are promising approaches.

Basically, there are two types of BCIs for EEG recordings: the invasive BCIs [1] and the non-invasive BCIs [10]. The non-invasive BCIs are used more widely in research and treatments due to their hurtlessness and safety. However, EEG signals acquired by the non-invasive BCI are more easily contaminated by other physiological signals such as EMG caused by facial muscle movements, especially when some patients are unable to control muscle movements because of their illnesses [3]. Figure 1 shows the influence of an EMG signal on clean EEG data. It can be seen that the clean EEG signal is almost destroyed by the EMG signal. The presence of noise interference presents formidable challenges in the realm of end-to-end EEG-based emotion recognition. Thus, it is important to design better end-to-end denoising neural networks.

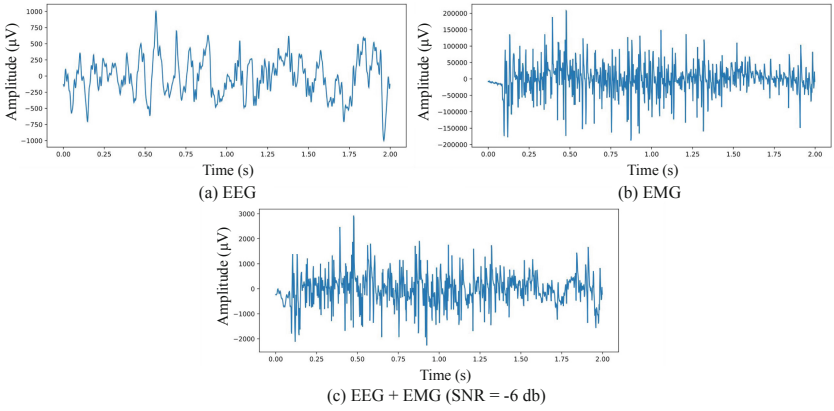


Fig. 1. Examples of a raw EEG segment and EMG interference

For end-to-end EEG-based emotion recognition, lots of endeavors have been made in the past few years. EEGnet [9] is a compact convolutional neural network (CNN) designed for EEG-based BCI to extract features from raw EEG signals.

To detect the valance and arousal levels, an end-to-end regional-asymmetric CNN was proposed and achieved an accuracy of over 95%, which was better than other methods using handcrafted features [16]. After that, Tao *et.al.* improved the accuracies to over 97% by an attention-based convolutional recurrent neural network (ACRNN) network on the same tasks [15]. These achievements show the superiority of end-to-end emotion recognition.

Deep learning methods have been proven to be effective in the denoising of EEG signals, which can learn the neural oscillations in EEG for eliminating noise from other artifacts. A benchmark dataset called EEGdenoiseNet was proposed for the research of EEG denoising [8]. However, it focuses on the EEG denoising task and not considering other EEG-based tasks.

To the best of our knowledge, addressing the processing of noisy data in the domain of end-to-end EEG-based emotion recognition remains largely unexplored. In this paper, we pioneerly introduce a novel Two-Stream Spectral-Temporal Denoising Network to achieve robust classification against EMG interference. An EEG noise-adding approach is proposed to simulate real-world muscle artifacts. Comprehensive experiments are conducted to test the proposed TS-STDN model in the cases of using clean data and EMG-contaminated data, respectively. Experimental results demonstrate the outperforming ability of our TS-STDN model in robust recognition. The code of our model and the noise-adding approach is published in https://github.com/XuanhaoLiu/TS-STDN.

2 Methodology

2.1 Data Preprocessing

Data Augmentation. The raw EEG signals with H Hz sampling frequency and duration T_{all} of a subject are denoted as $\mathbf{X}_{all} = [\mathbf{X}_{train}, \mathbf{X}_{test}] \in \mathbb{R}^{M \times C \times L}$, where M, C, and L is the number of trials, EEG channels, and sample points, respectively. One EEG trial $\mathbf{X}_s \in \mathbb{R}^{C \times L}$ is segmented into several slices $S = \{S_1, S_2, \ldots, S_n\}$ by sliding window, where slices $S_i (i = 1, 2, \ldots, n) \in \mathbb{R}^{C \times T}$. Due to the reason that the lengths of the EMG segments in EEGdenoiseNet are 2 s, we employ a 2-s sliding window with an overlap of 1-s for data augmentation. Hence, in this paper, we set $H = 200$ and $T = 400$.

Noisy Data Generation. The strategy for generating noisy data is to simulate real situations as much as possible. By carefully scanning a large amount of EEG records in the SEED dataset, we conclude that the brain areas which are prone to be disturbed by EMG signals are mainly distributed in the temporal areas on both sides of the scalp, while the frontal area and occipital area are less likely to be impacted. The influence of EMG signals is often asymmetric on the cortex as human muscle movements are not always symmetrical. Moreover, it cannot be ignored that individual differences are significant among different subjects. Based on the observation, forty electrodes that are easily affected by EMG signals are selected. The ten EMG groups are shown in Fig. 2, each group

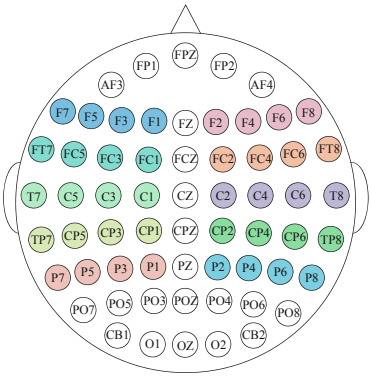


Fig. 2. The ten EMG noise-adding groups of electrodes, electrodes with the same color are in a group.

has four electrodes. We divided the forty electrodes into ten groups and for each particular EEG slice S_i, six out of ten groups are chosen randomly to be added the same EMG signals with a constant signal-noise ratio (SNR) calculated by Eq. (1), while there is no noise added to the remaining four groups. Notably, the EMG signals added to each EEG slice are selected randomly as well. The randomness creates asymmetry and individual differences in an easy way.

Let $x \in \mathbb{R}^T$ denotes a single channel of EEG signals, and $z \in \mathbb{R}^T$ denotes EMG signals, we generate noisy data by linearly combining x and λ times z with a constant SNR.

$$SNR = 10 \log \frac{RMS(x)}{RMS(\lambda \cdot z)}, \tag{1}$$

in which the RMS stands for root mean squared (RMS). Let $\mathbf{S} = \{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_n\}$, where slices $\mathbf{S}_i (i = 1, 2, \ldots, n) \in \mathbb{R}^{C \times T}$, denotes the noisy slices generated by the strategy we propose in this paper.

2.2 Two-Stream Spectral-Temporal Denoising Network

Inspired by the two-stream network for action recognition in videos [6], we propose a novel two-stream spectral-temporal denoising network with self-attention to fully excavate the emotional information from spectral and temporal aspects of EEG signals. Figure 3 illustrates the overall architecture of the two-stream spectral-temporal denoising network. The input noisy data are passed to both of the streams, which extract the temporal and spectral features, respectively. Meanwhile, two U-net networks [5] are employed to reconstruct the clean EEG signals in the temporal and spectral domains. After that, the features extracted by the CNN are concatenated and fed into an LSTM network [12]. Finally, all features are multiplied by the weight calculated with the self-attention module and classified by a linear layer to predict emotions.

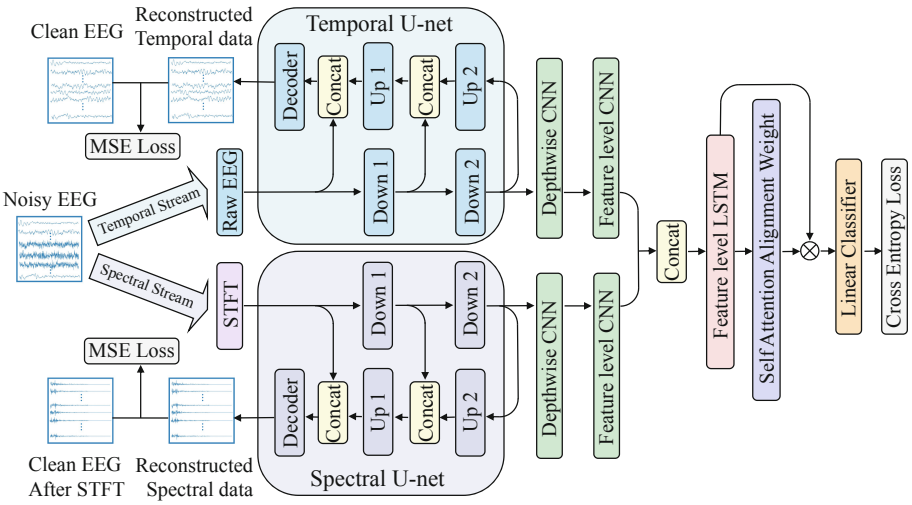


Fig. 3. Two-Stream Spectral-Temporal Denoising Network

Short Time Fourier Transform. To fully consider both the spectral and temporal information of EEG signals, short time Fourier transform (STFT) is utilized for extracting spectral information:

$$STFT(x,t) = \int_{-\infty}^{\infty} x(\tau)h(\tau - t)e^{-j2\pi f\tau}d\tau, \tag{2}$$

in which h is the window function. Let the $\tilde{S}' = \{\tilde{S}'_1, \tilde{S}'_2, \ldots, \tilde{S}'_n\}$ denotes the slices after STFT. For the purpose of balancing the size of two streams in our TS-STDN model, we elaborately choose the parameter of STFT to make the size of $\tilde{S}'$ approximately equal to raw EEG slices S.

Spectral and Temporal U-Net. By carefully setting the parameter of the STFT function, the slices $\tilde{S}$ are similar in size to S. Consequently, the spectral U-net and temporal U-net are designed to have the same structure symmetrically for balancing the stream scale and improving parallel performance. A batch of noisy EEG slices $\mathbf{S}_b \in \mathbb{R}^{B\times C\times T}$ is fed in the TS-STDN, where B is the size of each batch. During the training stage, the clean EEG batches S_b corresponding to the input noisy data $\mathbf{S}_b$ are used for training the denoising ability of TS-STDN. While at the testing stage, no clean data are available to our model.

The U-nets [5] can be regarded as encoder-decoder structures using several down or up units with shortcut concatenation, where each unit consists of a 1-D convolutional layer and an average pooling layer or an upsampling layer. The convolutional layers keep the length of the input and output tensors consistent by padding but change the channel numbers simultaneously. The lengths of tensors are doubled up by the upsampling layer, or reduced by half by the average pooling layer with a kernel of size (1,2).

We only introduce the temporal U-net due to the symmetry. Firstly, the input batch $\mathbf{S}_b$ is transformed to $\mathbf{S}'_b \in \mathbb{R}^{B\times 1\times C\times T}$ for subsequent calculating. Let $\sigma()$ denote the exponential linear unit (ELU) activation function $ELU()$, and $BN()$ stands for the batch normalization. Consequently, the outputs of down1 and down2 are:

$$X_{d1} = AvgPool(\sigma(BN(Conv_{d1}(\mathbf{S}'_b)))) \in \mathbb{R}^{B\times 2\times C\times (T/2)}, \tag{3}$$

$$X_{d2} = AvgPool(\sigma(BN(Conv_{d2}(X_{d1})))) \in \mathbb{R}^{B\times 4\times C\times (T/4)}. \tag{4}$$

The up1 and up2 units are designed to reduce half of the channel number but double the length of the input tensor by replacing the average pooling layer with an upsampling layer in the down units. Consequently, the outputs of up1 and up2 are calculated as:

$$X_{u2} = \sigma(BN(Conv_{u2}(Upsample(X_{d2})))) \in \mathbb{R}^{B\times 2\times C\times (T/2)}, \tag{5}$$

$$X_{u1} = \sigma(BN(Conv_{u1}(Upsample(Concat(X_{u2}, X_{d1}))))) \in \mathbb{R}^{B\times 2\times C\times T}. \tag{6}$$

Afterward, the output of up1 X_{u1} concatenated with the input noisy data $\mathbf{S}'_b$ is fed into the decoder, whose output has the same shape of tensor $\mathbf{S}'_b$. The decoder has two convolutional layers to better reconstruct the clean EEG signals without EMG interference $\hat{S}_b$. The reconstructed data $\hat{S}_b$ is used to compute the Mean-Squared Loss (MSE) loss with the clean data S_b:

$$\mathcal{L}_{temporal} = MSELoss(\hat{S}_b, S_b). \tag{7}$$

Same to the reconstruction loss in the temporal domain, the reconstruction loss of spectral signals is calculated by:

$$\mathcal{L}_{spectral} = MSELoss(\hat{\tilde{S}}_b, \tilde{S}_b). \tag{8}$$

Depthwise CNN and Feature Level CNN & RNN. Inspired by the Xception [7], a depthwise CNN with a kernel size of $(C, 1)$ is employed for aggregating the spatial information between EEG channels. Let D denotes the depth multiplier number, the output of the depthwise CNN is:

$$X_D = AvgPool(\sigma(BN(Conv_D(X_{d2})))) \in \mathbb{R}^{B\times 4D\times 1\times (T/8)}. \tag{9}$$

The feature level CNN is utilized to compress temporal information. Let F_t and F_s denote the filter numbers of the feature level CNN in the temporal and spectral streams, the output of the feature level CNN is:

$$X_{F_{s/t}} = AvgPool(\sigma(BN(Conv_F(\tilde{X}_D/X_D)))) \in \mathbb{R}^{B\times F_{s/t}\times 1\times (T/32)}. \tag{10}$$

An LSTM network is adopted for the spectral and temporal information fusion. The output of the feature level CNN of both streams X_{F_s} and X_{F_t} are reshaped, concatenated and linearly embedded to $X_F \in \mathbb{R}^{B\times (F_s+F_t)\times d}$, where d is the embedding dimension. As the LSTM is a seq2seq model, the X_F is transformed to $X_L \in \mathbb{R}^{B\times (F_s+F_t)\times d}$ by regarding each feature as a token of d dimension.

Self-Attention Alignment Weight. After extracting the spatial, spectral, and temporal information through two-stream U-net, CNN, and LSTM, the output tensor X_L is highly semantic and discriminative. However, some of the $F_a = F_s + F_t$ features are not closely related to human emotions. Hence, a self-attention module [4] is utilized for concentrating on the features mostly related to emotions. The alignment weight is calculated from the average of each feature with two linear layers activated by the $tanh()$ function. Let $X_a \in \mathbb{R}^{B \times F_a}$ denotes the average tensor of X_L. The first dimensionality-reduction layer has a parameter $W_1 \in \mathbb{R}^{F_a \times d_r}$ and a bias $b_1 \in \mathbb{R}^{d_r}$, while the second dimensionality increasing layer has a parameter $W_2 \in \mathbb{R}^{d_r \times F_a}$ and a bias $b_2 \in \mathbb{R}^{F_a}$, where the d_r is the reduction dimension number. The alignment weight w is calculated by:

$$w = softmax(W_2 \cdot (tanh(W_1 \cdot X_a + b_1)) + b_2) \in \mathbb{R}^{B \times F_a}. \tag{11}$$

Finally, multiply each feature of X_L by the corresponding coefficient in the attention weights w to produce X_{sa}. X_{sa} is $\{w_1 x_1, w_2 x_2, w_3 x_3, \ldots, w_{F_a} x_{F_a}\}$. Then, X_{sa} is flattened and fed into a linear classifier to get the final prediction $\hat{y}$. The cross-entropy loss is applied to compute the classification loss:

$$\mathcal{L}_{cls} = -\sum_{i=1}^{N} y_i \log \hat{y}_i \ , \tag{12}$$

where y stands for the ground truth emotion label. Consequently, the entire objective function is given as minimizing the linear combination of the classification loss and reconstruction loss:

$$\arg\min_{\Theta} \mathcal{L}_{all} = \arg\min_{\Theta} (\mathcal{L}_{cls} + \mathcal{L}_{spectral} + \mathcal{L}_{temporal}). \tag{13}$$

3 Experiment

3.1 Dataset

SEED and SEED-IV. The SJTU Emotion EEG datasets are a series of datasets that record the EEG signals of subjects while they are watching emotion videos. The original SEED dataset [10] chooses fifteen Chinese film clips in order to induce three target emotions: positive, neutral, and negative. The SEED-IV dataset [11] contains four categories of emotion including happy, sad, neutral, and fear. Seventy-two film clips are chosen as stimuli. The 62-channel ESI NeuroScan System was employed to capture EEG signals in both datasets, using the EEG cap with 62 channels positioned according to the international 10–20 system at 1000 Hz. All EEG signals are then downsampled to 200 Hz.

Specifically, for the SEED dataset, each recording contains fifteen trials of EEG signals, we use the samples from the first nine clips as the training set and the samples from the remaining six clips as the testing set. However, for the SEED-IV dataset, the last eight clips are unequal in the number of emotion types. Hence, we choose the two video clips that appear at the end of each session for each emotion as the testing set, and the remaining sixteen clips compose the training set.

EEGdenoiseNet. EEGdenoiseNet [8] is a benchmark dataset designed for the purpose of training and evaluating deep learning denoising models. The dataset consists of single-channel EEG, EOG, and EMG signals collected from diverse publicly available datasets. To ensure the quality and reliability of the dataset, rigorous preprocessing procedures are conducted on all physiological signals. Notably, the signals are segmented into 2-second intervals and meticulously examined by an expert to confirm their cleanliness and suitability for analysis. There are 5598 pure EMG segments in EEGdenoiseNet, and as the original EMG signals are 512 Hz, we downsampled them to 200 Hz to match the EEG signals from SEED datasets.

3.2 Implementation Details

We compare our TS-STDN model with other end-to-end approaches including LSTM [12], EEGnet [9], and ACRNN [15] under subject-dependent conditions. In this paper, the number of EEG channels is $C = 62$, the sampling frequency is $H = 200$ Hz, and the length of slices is $T = 400$. For the TS-STDN model, the depth multiplier number $D = 4$, the filter number of the feature level CNN of both streams are $F_s = F_t = 16$, and the embedding dimension $d = 16$. The LSTM model regards the input EEG data as a sequence that contains T tokens, each token is obtained by embedding a sample point in a 32-dim tensor through two linear layers connected by a ReLU activation function. The hyperparameters of the EEGnet model and ACRNN remain the same as those in the original papers. All models are implemented by PyTorch [13] deep learning framework and trained with Adam optimizer with a learning rate of $\eta = 0.001$ and a batch size of $B = 64$.

3.3 Ablation Experiments

Ablation experiments are conducted by removing some modules to evaluate the effectiveness of each module in our TS-STDN model. We design several variant models as follows:

- Spectral DN: This model only contains the spectral stream of the TS-STDN.
- Temporal DN: This model only contains the temporal stream of the TS-STDN.
- TS-STDN w/o SA: This model only removes the self-attention alignment weight module.
- TS-STDN w/o MSE: This model is trained without using the reconstruction MSE Loss computed with clean data in both spectral and temporal domains.

3.4 Results on Clean Data

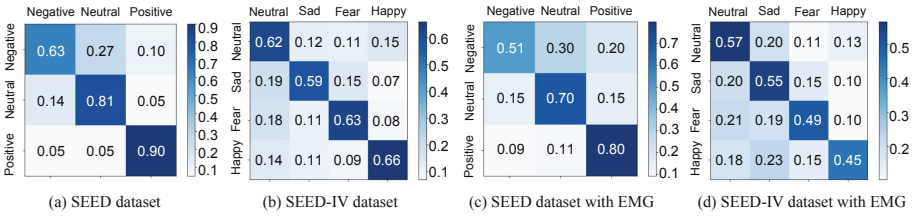
We use the raw EEG signals without adding noise to evaluate the performance of each model. Subject-dependent experiments on two public datasets SEED

Table 1. The accuracies (Avg./Std.) of different methods on SEED and SEED-IV

Method	SEED		SEED-IV	
	Avg. (%)	Std. (%)	Avg. (%)	Std. (%)
LSTM [12]	67.50	12.06	43.95	10.30
ACRNN [15]	57.37	10.93	40.19	07.27
EEGnet [9]	72.49	12.52	54.26	12.19
Spectral DN	76.88	09.89	54.90	11.01
Temporal DN	73.69	12.70	55.50	12.17
TS-STDN w/o SA	75.73	09.96	56.64	10.12
TS-STDN w/o MSE	76.93	10.98	58.84	12.11
TS-STDN	**78.45**	10.49	**62.13**	12.18

[10] and SEED-IV [11] demonstrates the outperforming performance of our TS-STDN model, the classification accuracies are presented in Table 1. Remarkably, our TS-STDN achieves the highest accuracies of 78.45% and 62.13% on the SEED and SEED-IV datasets, respectively. It is worth noting that relying solely on a single stream results in varying decreases in accuracy, which shows the complementary properties of the temporal and spectral information. The Spectral DN demonstrates better performance on the SEED dataset, whereas the Temporal DN exhibits higher accuracy on the SEED-IV dataset. As a result, it remains uncertain which stream holds greater importance over the other. Without employing the self-attention module, TS-STDN fails to extract the key features closely related to emotions and has 3% and 6% reductions in accuracy. It can be seen that the denoising U-net is still effective when processing clean data, which can be regarded as an autoencoder.

The confusion matrices illustrating the classification performance of our TS-STDN model on the SEED and SEED-IV datasets are presented in Fig. 4(a) and (b). Regarding the SEED dataset, the positive emotion exhibits the highest accuracy, whereas classifying the negative emotion proves to be notably challenging. In contrast, for the SEED-IV dataset, the accuracies for all emotions are relatively similar to each other.



(a) SEED dataset (b) SEED-IV dataset (c) SEED dataset with EMG (d) SEED-IV dataset with EMG

Fig. 4. Confusion matrices of our TS-STDN model

Table 2. The average accuracies (%) of different methods on the SEED dataset under varying levels of EMG interference.

Method	The SNR of contaminated EEG						Avg
	−6 db	−7 db	−8 db	−9 db	−10 db	−11 db	
LSTM [12]	51.95	55.22	55.02	48.57	48.74	48.57	51.34
ACRNN [15]	48.30	49.68	49.30	47.05	47.43	46.90	48.11
EEGnet [9]	57.32	59.63	59.08	54.52	53.67	54.19	56.40
Spectral DN	65.51	68.26	68.10	63.23	62.58	63.83	65.25
Temporal DN	64.79	67.89	66.77	61.29	61.24	63.05	64.17
TS-STDN w/o SA	64.62	68.58	68.94	64.08	63.64	64.05	65.65
TS-STDN w/o MSE	65.50	68.36	68.72	61.26	61.00	61.03	64.31
TS-STDN	**67.81**	**70.38**	**71.03**	**65.27**	**64.79**	**64.67**	**67.33**

3.5 Results on Noisy Data

To fully investigate the robustness of our TS-STDN model against the interference of EMG signals, we tested our model under conditions where the raw EEG signals are contaminated by EMG signals with varying signal-to-noise ratios. The clean EEG data are from the SEED dataset [10] and the SEED-IV dataset [11], while the EMG data are from the EEGdenoiseNet dataset [8]. Specifically, we conducted our experiment on noisy data with SNR from −6 dB to −11 dB, simulating the disturbance from relatively moderate to extremely intense. The SNR is calculated by the Eq. (1). For fairness, all methods are tested on the same noisy data which had been generated before evaluation.

Table 2 and Table 3 display the results of each model under various noise intensity conditions on the noisy SEED and SEED-IV datasets, respectively. It can be seen that our TS-STDN model acquires the best classification accuracies under all conditions, exhibiting an increase of about 11% to the EEGnet on the SEED dataset, and 10% on the SEED-IV dataset to the EEGnet. All baseline models, which possess no denoising modules, perform worse when processing contaminated EEG. The Spectral DN model shows better performance than the Temporal DN model when facing noisy data on both datasets, suggesting that spectral information is more robust to EMG interference. Through the observation of reduced accuracy on both datasets of the TS-STDN w/o SA model, it is evident that the self-attention module remains effective when facing noisy EEG data. It is worth noting that by employing reconstruction loss, our TS-STDN model exhibits a notable improvement in accuracy, demonstrating an increase of 3% on the SEED dataset and 7% on the SEED-IV dataset to the TS-STDN w/o MSE model, which proves the effectiveness of the reconstruction module. Moreover, compared to the improvement of the reconstruction module on the clean data, which are 2% and 4% on two datasets, we find the reconstruction module is more suitable for denoising. The average confusion matrices of our

TS-STDN model on noisy datasets are depicted in Fig. 4(c) and (d). All emotion classification accuracies are influenced by EMG disturbance.

Table 3. The average accuracies (%) of different methods on the SEED-IV dataset under varying levels of EMG interference.

Method	The SNR of contaminated EEG						Avg
	−6 db	−7 db	−8 db	−9 db	−10 db	−11 db	
LSTM [12]	33.64	33.33	33.63	33.61	33.48	33.22	33.48
ACRNN [15]	37.40	37.67	37.36	37.72	37.35	37.00	37.42
EEGnet [9]	42.57	42.86	42.41	42.83	42.07	41.98	42.46
Spectral DN	48.93	49.11	48.21	49.03	49.51	48.73	48.92
Temporal DN	46.21	45.13	45.22	45.34	46.02	45.72	45.61
TS-STDN w/o SA	50.12	49.15	49.15	49.34	50.31	49.66	49.62
TS-STDN w/o MSE	45.09	44.80	45.73	44.85	45.79	45.07	45.22
TS-STDN	**52.17**	**52.40**	**52.84**	**52.25**	**52.33**	**51.92**	**52.32**

4 Conclusion

In this paper, we introduce a novel Two-Stream Spectral-Temporal Denoising Network to thoroughly exploit emotion-related features from both spectral and temporal views in contaminated EEG signals while learning the ability to eliminate noise interference. The TS-STDN model acquires the denoising capability by reconstructing clean EEG signals using U-net architectures, which can be seen as a denoising autoencoder. To simulate the real-world EMG disturbance on EEG signals, we propose a random algorithm for EMG noise adding in the raw EEG recordings. The experimental results demonstrate that the TS-STDN model performs the best on both clean and contaminated data, and is robust when facing extremely intense noise interference. The source code of the noisy data production and TS-STDN model implementation is shared in https://github.com/XuanhaoLiu/TS-STDN.

Acknowledgments. This work was supported in part by grants from National Natural Science Foundation of China (Grant No. 61976135), STI 2030-Major Projects+2022ZD0208500, Shanghai Municipal Science and Technology Major Project (Grant No. 2021SHZDZX), Shanghai Pujiang Program (Grant No. 22PJ1408600), Medical-Engineering Interdisciplinary Research Foundation of Shanghai Jiao Tong University "Jiao Tong Star" Program (YG2023ZD25), and GuangCi Professorship Program of RuiJin Hospital Shanghai Jiao Tong University School of Medicine.

References

1. Lorach, H., Galvez, A., Spagnolo V., Martel, F., et al.: Walking naturally after spinal cord injury using a brain-spine interface. Nature 1–8 (2023)
2. Alarcao, S.M., Fonseca, M.J.: Emotions recognition using EEG signals: a survey. IEEE Trans. Affect. Comput. **10**(3), 374–393 (2017)
3. Supriya, S., Siuly, S., Wang, H., Zhang, Y.: Epilepsy detection from EEG using complex network techniques: a review. IEEE Rev. Biomed. Eng. **16**, 292–306 (2021)
4. Vaswani, A., Shazeer, N., Parmar, N.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008 (2017)
5. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
6. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 568–576 (2014)
7. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1251–1258 (2017)
8. Zhang, H., Zhao, M., Wei, C., Mantini, D., Li, Z., Liu, Q.: EEGdenoiseNet: a benchmark dataset for deep learning solutions of EEG denoising. J. Neural Eng. **18**(5), 056057 (2021)
9. Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J.: EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. J. Neural Eng. **15**(5), 056013 (2018)
10. Zheng, W.L., Lu, B.L.: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Trans. Auton. Ment. Dev. **7**(3), 162–175 (2015)
11. Zheng, W.L., Liu, W., Lu, Y., Lu, B.L., Cichocki, A.: Emotionmeter: a multimodal framework for recognizing human emotions. IEEE Trans. Cybern. **49**(3), 1110–1122 (2018)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
13. Paszke, A., et al.: Automatic differentiation in pytorch (2017)
14. Duan, R.N., Zhu, J.Y., Lu, B.L.: Differential entropy feature for EEG-based emotion classification. In: 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), pp. 81–84. IEEE (2013)
15. Tao, W., et al.: EEG-based emotion recognition via channel-wise attention and self attention. IEEE Trans. Affect. Comput. **14**(1), 382–393 (2023)
16. Cui, H., Liu, A., Zhang, X., Chen, X., Wang, K., Chen, X.: EEG-based emotion recognition using an end-to-end regional-asymmetric convolutional neural network. Knowl.-Based Syst. **205**, 106243 (2020)
17. Li, R., Wang, Y., Zheng, W.L., Lu, B.L.: A multi-view spectral-spatial-temporal masked autoencoder for decoding emotions with self-supervised learning. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 6–14 (2022)